Master 2 Data Science
TD1 - Cours de Deep Learning
MLP et rétropropagation du gradient

I. Optimisation d'un réseau de neurones sans couche cachée

On considère un réseau de neurones sans couche cachée à 1 sortie pour un problème de classification biclasses (un exemple quelconque appartient à une classe et une seule parmi les 2 classes données). Ce réseau a donc une sortie et calcule pour une entrée (vectorielle) x une quantité $F_w(x) = g(w^T x)$, où g est une fonction d'activation linéaire ou non linéaire.

L'exemple est classé dans la classe 1 ou 2 suivant que $F_w(x)$ est plus proche de la cible correspondant à la classe 1 ou de celle correspondant à la classe 2 (on utilise généralement les cibles $\{0, 1\}$ ou $\{-1, 1\}$.

L'apprentissage du modèle est réalisé en minimisant le risque empirique sur un ensemble d'apprentissage $\{(x^i, y^i), i = 1...N\}$, en utilisant un critère d'erreur quadratique.

(1) Quelle est la forme de la frontière de décision entre les deux classes implementée par ce modèle ? La nature de cette frontière dépend-t-elle de la fonction d'activation g ?

(2) On utilise une fonction d'activation g linéaire et des cibles $\{-1, 1\}$. Dérivez le critère d'optimisation pour déterminer le gradient. Quels sont les points pour lesquels le gradient est le plus fort en amplitude ? Sont-ils les points les moins bien classés du jeu d'entrainement ? Illustrez votre raisonement sur des données (séparables linéairement) en dimension 2.

(3) Mêmes questions pour le cas où g est non linéaire saturante (par exemple *tanh*).

II. Approximation des probabilités a posteriori par un réseau de neurones appris avec un critère MSE sur des cibles 0/1

On considère un réseau de neurones à 1 sortie pour un problème de classification biclasses (un exemple quelconque appartient à une classe et une seule parmi les 2 classes données).

(1) On note $C(w)$ le risque réél, où l'on utilise un critère d'erreur quadratique. Ecrivez comment ce risque s'exprime.

(2) Montrez que la solution optimale optimisant ce risque est que le réseau de neurones calcule $F_w(x) = P(y = 1|x)$

(3) Généralisez la démonstration au cas à K classes.

(4) La démonstration précédente est elle particulière aux réseaux de neurones ?

III. Critère d'optimisation: Erreur quadratique, Cross entropy, Weight Decay

On considère un réseau de neurones à K sorties pour un problème de classification multiclasses (un exemple quelconque appartient à une classe et une seule parmi les K classes données).

(1) Pour un exemple d'apprentissage (x, y) où $y \in \{0, ..., K\}$, le critère d'erreur quadratique consiste à calculer la norme de la différence entre le vecteur produit en sortie par le réseau (pour x mis en entrée), $F_w(x)$, et du one hot vecteur indicateur de la classe, vecteur à K dimension également avec une composante à 1 (la y^{ieme}) et les autres à 0. Ecrivez ce critère puis la dérivée de ce critère pour un exemple d'apprentissage, par rapport aux sorties du réseau de neurone dans le cas où les neurones de sortie ont des fonctions d'activation linéaires.

(2) Souvent pour des problèmes de classification, on utilise une fonction d'activation *softmax* sur la couche de sortie. En notant $(s_k, k = 1..K)$ les activations de la dernière couche du réseau, avant la couche d'activation softmax, la fonction d'activation *softmax* calcule de nouvelles sorties égales à $f_w^k(x) = \frac{e^{s_k}}{\sum_{j=1...K} e^{s_j}}$. Quel est l'intérêt de cette fonction d'activation selon vous ?

(3) Lorsque l'on utilise des activations softmax, on utilise souvent le critère d'entropie croisée qui est défini comme suit. Pour un exemple d'apprentissage (x, y) où $y \in \{0, ..., K\}$, le critère d'entropie est défini par: $-\sum_{k=1...K} t_k \ln f_w^k(x)$ où $f_w(x)$ représente le vecteur de sorties du réseau de neurones (de dimension K) et où $f_w^k(x)$ représente sa k^{ime} composante et où t_k est égal à 1 si $y = k$ et 0 sinon. Ecrivez la dérivée de ce critère pour un exemple d'apprentissage, par rapport aux sorties du réseau de neurones. En interprétant la sortie $f_w^k(x)$ comme la probabilité a posteriori $p(y = k|x)$ montrer que minimiser le critère d'entropie croisée correspond a maximiser la vraisemblance conditionnelle des données.

(4) La stratégie de Weight Decay, autrement dit une régularisation L2, consiste à ajouter un terme dans la fonction objectif égal à $\|w\|^2$, c'est à dire la norme des poids, mis sous la forme d'un seul vecteur. La fonction objectif devient ansi $O(w) = C(w) + \lambda\|w\|^2$. En notant w_{ij}^l le poid du neurone j de la couche $l - 1$ vers le neurone i de la couche l, qu'est-ce que l'ajout du terme de Weight Decay change sur le gradient de la fonction objectif par rapport à un poid du réseau w_{ij}^l ?